Statistics Report

Master Data Science & Engineering - FEUP

Fundamentos de Ciência e Engenharia dos Dados

12/11/2021

Grupo 6

Fábio Fernandes e Hugo Nogueira

Study Methodology

Six-Steps Statistical Investigation

Method used on this study

Research question

What difference is there in the 2 districts with the greatest forest fire impact?

Study & collect data

01 Study

forest fires in Portugal IN 2015.
we analyze and filter as variables of interest Portugal has been plagued by forest fires and as a study will be carried out, we will assess their characteristics, relationships and meanings.

02 Dataset

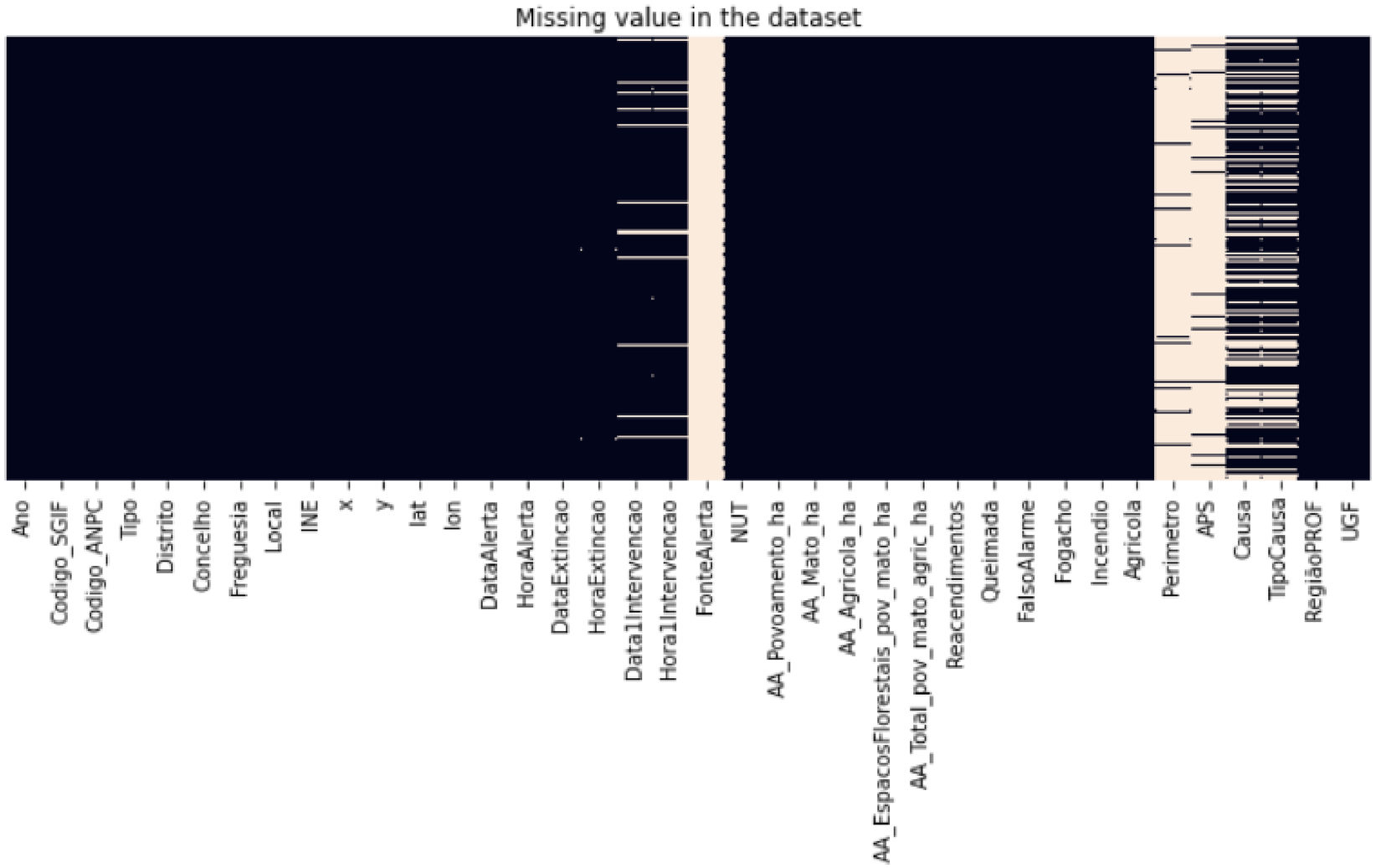
Used data was collected by
Rural Fires 2025
Instituto de conservação da Natureza e das Florestas.
DECIF - Dispositivo Especial de Combate a Incêndios Florestais

Exploratory data

Dataset Exploratory

```
#dataset size
print(df.shape)
```

```
(23175, 38)
```

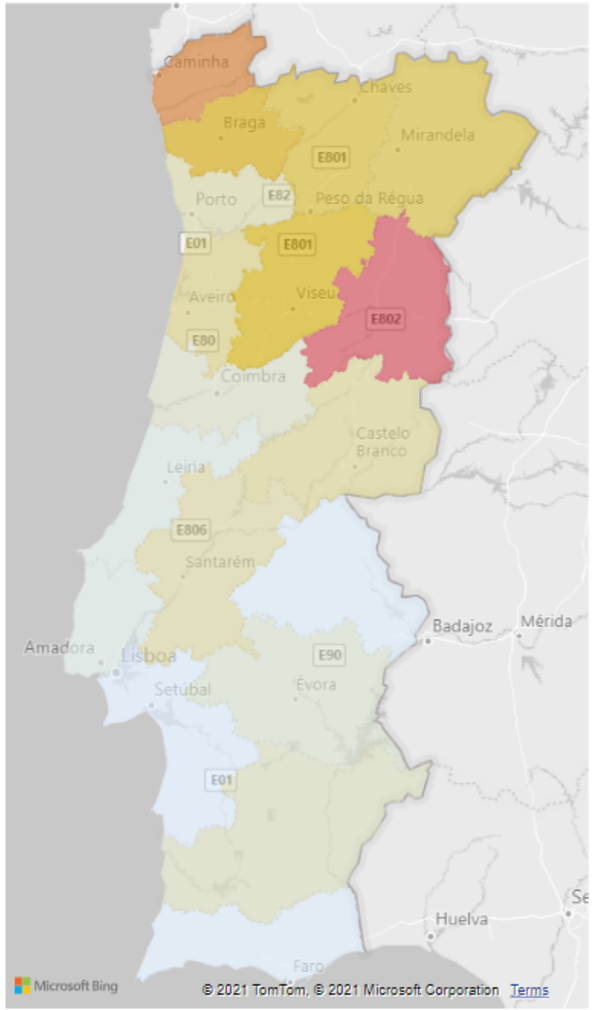


Exploratory data

Dataset Exploratory

18,90% of burned area in Guarda
14,70% of burned area Viana do Castelo
out of total 68k acres.

Distrito



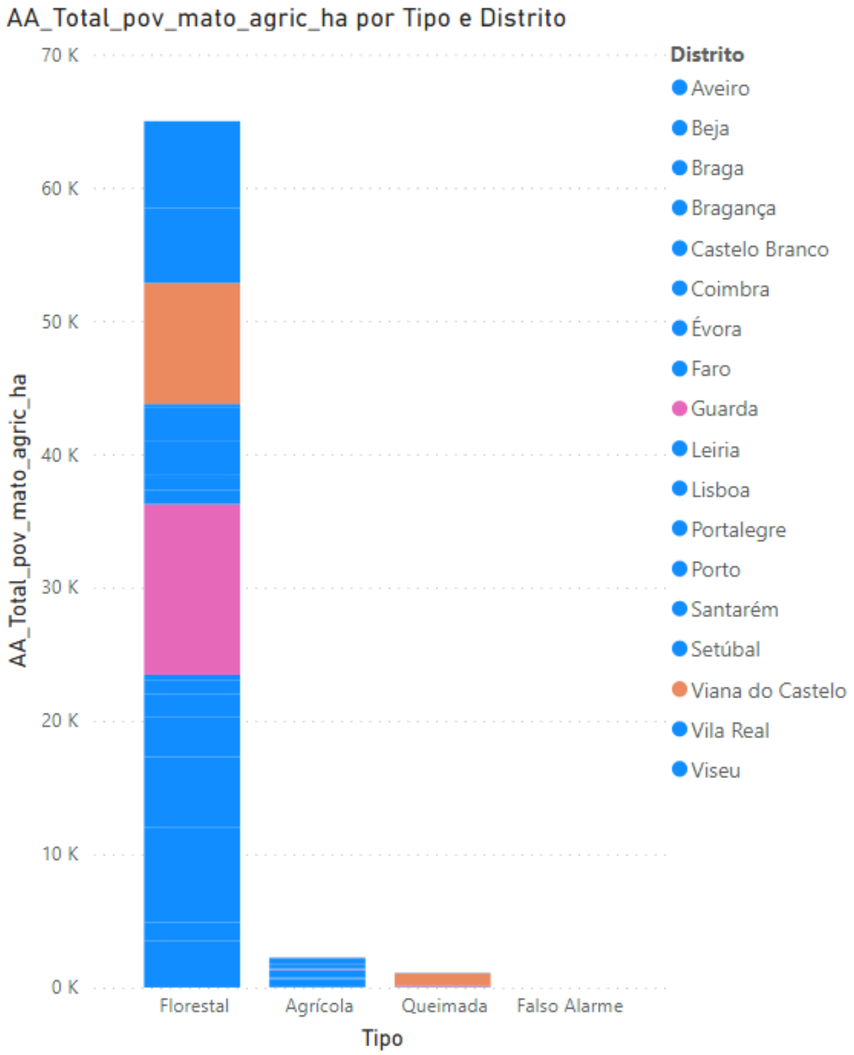
AA_Total_pov_mato_agric_ha	%GT AA_Total_pov_mato_agric_ha	Number_fires	Distrito
12.914,63	18,90%	809	Guarda
10.044,43	14,70%	1559	Viana do Castelo
7.178,27	10,51%	2231	Braga
6.572,86	9,62%	1528	Viseu
5.618,64	8,22%	1481	Vila Real
5.299,58	7,76%	733	Bragança
3.521,48	5,15%	1774	Aveiro
3.113,27	4,56%	544	Castelo Branco
2.708,73	3,96%	1328	Santarém
2.545,86	3,73%	4444	Porto
1.937,71	2,84%	438	Beja
1.739,14	2,55%	749	Coimbra
1.568,84	2,30%	306	Évora
1.079,68	1,58%	2225	Lisboa
1.078,11	1,58%	906	Leiria
508,08	0,74%	363	Portalegre
478,53	0,70%	584	Faro
423,77	0,62%	1173	Setúbal
68.331,59	**100,00%**	**23175**	

Exploratory data

Dataset Exploratory

95% of burned area tipo Florestal out of 68k acre.

Tipo	AA_Total_pov_mato_agric_ha	%GT AA_Total_pov_mato_agric_ha
Florestal	64.977,98	95,09%
Agrícola	2.256,18	3,30%
Queimada	1.097,43	1,61%
Falso Alarme	0,00	0,00%
Total	**68.331,59**	**100,00%**



Exploratory data

Dataset

set of features including:

Category:

* `tipo`: the type of fire. **'Floresta**
* `Distrito`: District / location of the fire - **Guarda e Viana do Castelo**
* `DataALertas`: Date and Time of alert.
* `Data1Intervencao`: Date and time of firsrt intervenction
* `DataExtincao`: Date and time of extinction.

Numeric:

* `areaArdia`: Burned area.

New features:

* `Tempo_de_resposta`: **Response time** . datediff(Data1Intervenção - dataAlertas)
* `tempo_de_extincao`: **Extinct date and time.** datediff(DataExtincão - Data1Intervencao)

Exploratory data

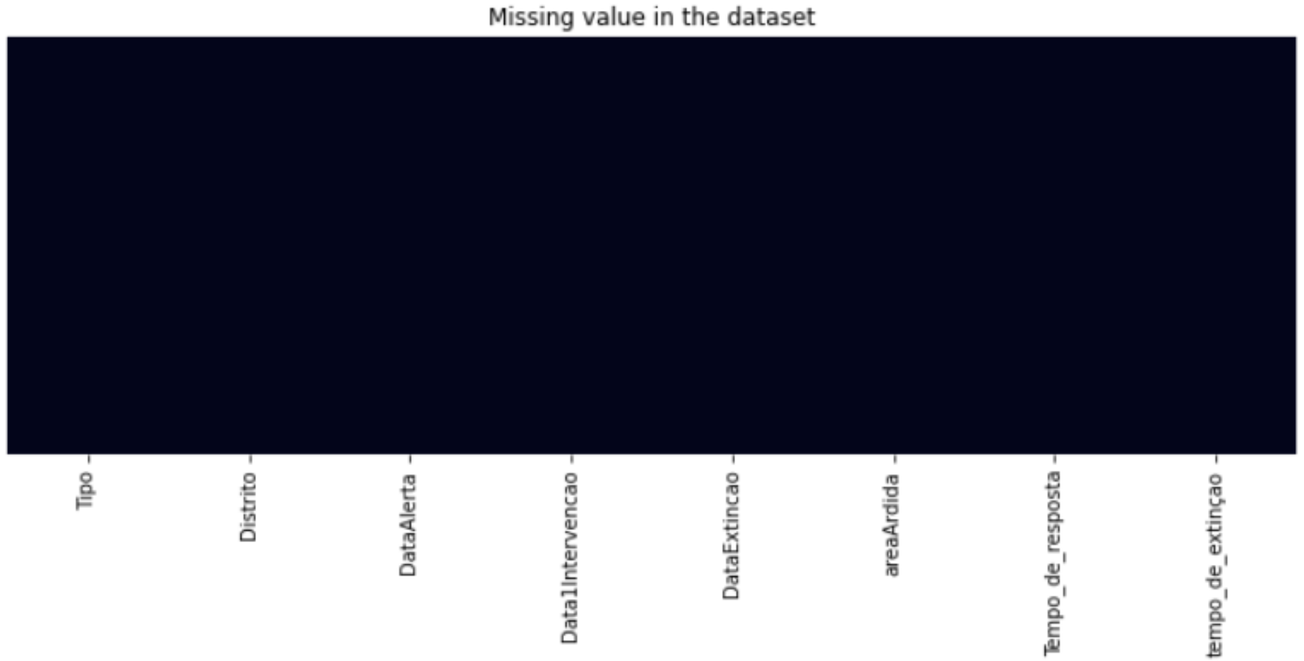
Dataset Exploratory

Dataset statistics

Number of variables	9
Number of observations	1545
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	597.8 KiB
Average record size in memory	396.2 B

Variable types

Numeric	4
Categorical	5



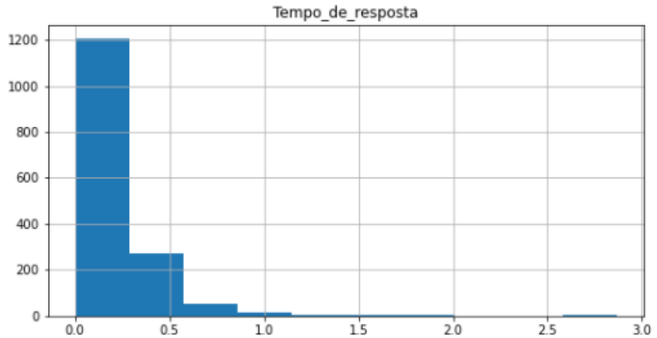
Exploratory data

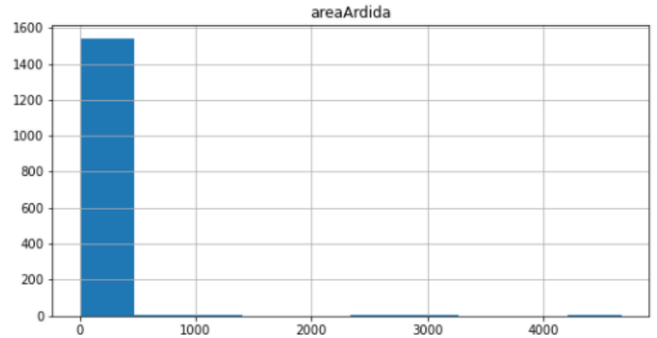
- Descriptive statistics

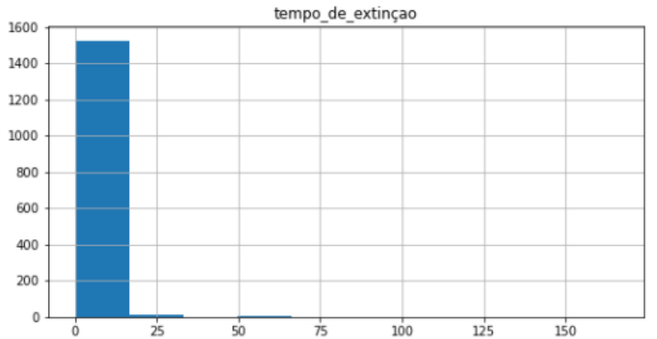
Summary Statistics

	areaArdida	Tempo_de_resposta	tempo_de_extinçao
count	1545.00	1545.00	1545.00
mean	14.00	0.21	2.90
std	166.40	0.20	7.70
min	0.00	0.00	0.02
25%	0.02	0.08	1.08
50%	0.20	0.18	1.73
75%	1.50	0.27	2.83
max	4673.00	2.87	165.63

Histogram







Exploratory data

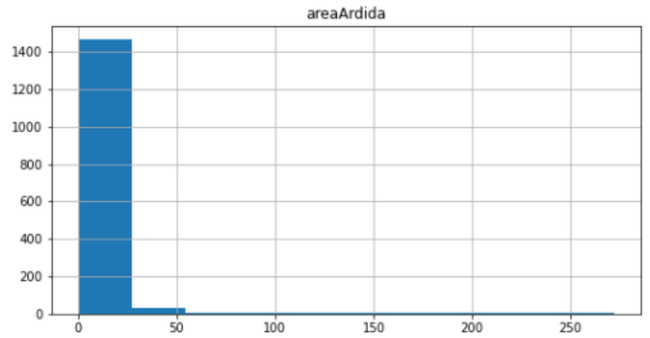
- Descriptive statistics - Zscore (eliminate Outliers)

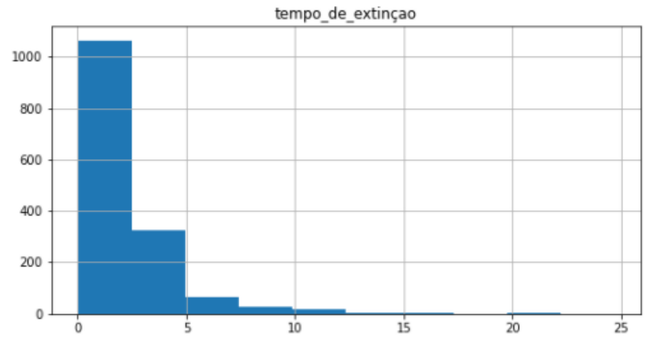
Tells how many standard deviations away a given observation is from the mean, config > 3 SD

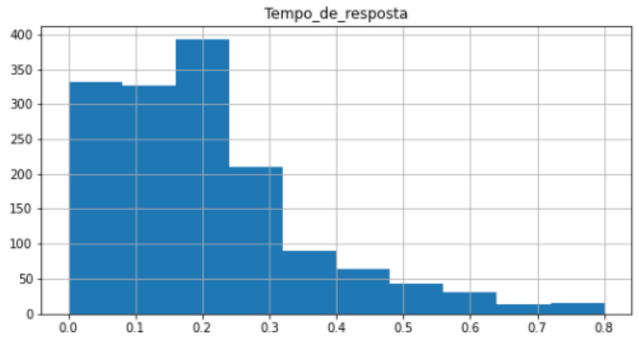
Summary Statistics

	areaArdida	Tempo_de_resposta	tempo_de_extinçao
count	1513.00	1513.00	1513.00
mean	3.92	0.20	2.36
std	18.22	0.15	2.32
min	0.00	0.00	0.02
25%	0.02	0.08	1.08
50%	0.20	0.17	1.72
75%	1.50	0.27	2.80
max	272.00	0.80	24.67

Histogram







Inferences

Significance Estimation

As we are interested in understanding the impact of the fires in the forest, we want to find it, if there's significant differences between the Burned Ares of the two locations and if there is some relation between the amount of Burned Area and the Response Time to the fires.

Thus,

Question 1:

- **Is the mean Burned Area per fire significantly different in the two locations?**

    Remeber the means: Guarda = 6.06 and Viana do Castelo = 3.05 ha

Question 2:

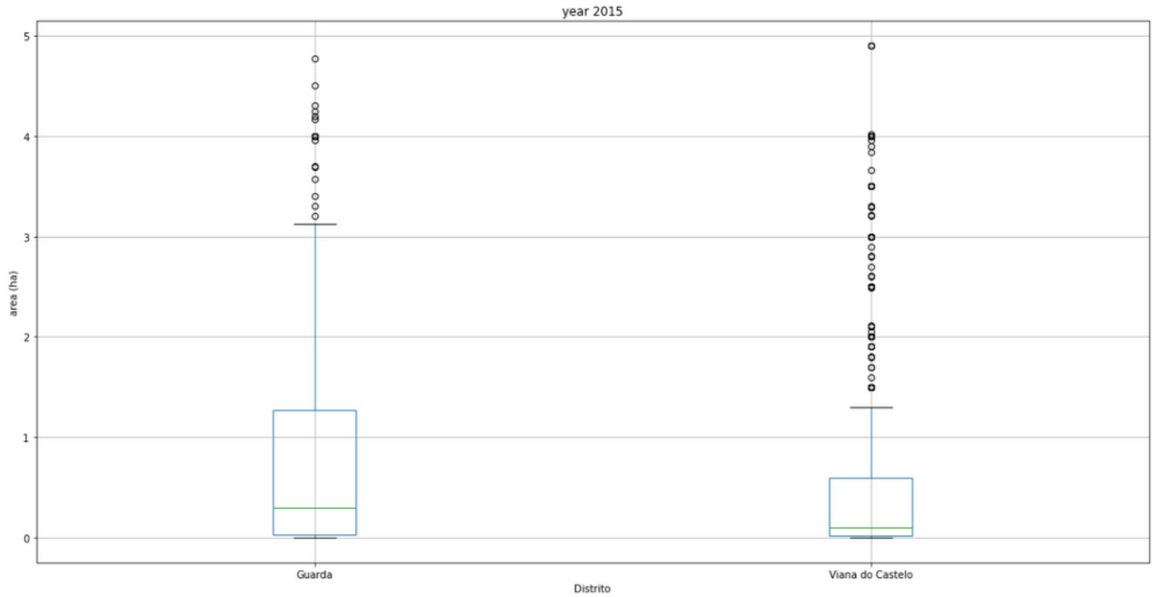
- **Is there a relation between the Burned Area and the time to Response that causes more or less impact in the forest fires?**

Inferences

Question 1:

- Is the mean Burned Area per fire significantly different in the two locations?

Distrito	count	mean	std	min	25%	50%	75%	max
Guarda	436.0	6.068574	23.014143	0.00005	0.05	0.50	3.0	250.0
Viana do Castelo	1077.0	3.052255	15.799523	0.00010	0.02	0.12	1.0	272.0



For burned area (BA) we have:

- Guarda with a mean BA by fire of 6.06 acres

- Viana do Castelo with a mean BA of 3.05

It looks that Guarda has more BA per fire than Vlana do Castelo, although Viana do Castelo has more fire frequency

Inferences

Question 1:
- Is the mean Burned Area per fire significantly different in the two locations?

Two sample t-test with equal variances

Guarda variance: 528.435
Viana do Castelo variance: 249.393
variance ratio: 2.118

2.11 < 4 This means we can assume that the population variances are equal. Thus, we can proceed to perform the two sample t-test with equal variances

Thus, we can proceed to perform the two sample t-test with equal variances

Note:
- Not checked if data in each group are normally distributed.
- Assumed that 2015 data is a sample from the entire fires population from all Portugal locations and data for each location independent

Inferences

Question 1:

- Is the mean Burned Area per fire significantly different in the two locations?

The **t test statistic is 2.924** and the corresponding **two-sided p-value is 0.003.**

The two hypotheses for this particular two sample t-test are as follows:

H0: μ1 = μ2 (the two population means are equal)

HA: μ1 ≠μ2 (the two population means are not equal)

Because the p-value of our test (0.003) is less than alpha = 0.05,

We r**eject the null hypothesis of the test.**

Acepting the alternative hypothesis says that **the two population means are not equal or, the mean Burned area from this two locations are not equal**

Inferences

Question 2:

- Is there a relation between the Burned Area and the time to Response that causes more or less impact?

Independence test for continuous variables

In **general, Pearson correlation ρ does not represent the independence or dependence,** but a linear relationship between two random variables.

If we assume that paired two random variables are either independent or just linearly related, then Pearson correlation ρ can be used to measure independency.

The **covariance and covariance matrix** are used widely within statistics and multivariate analysis to characterize the relationships between two or more variables.

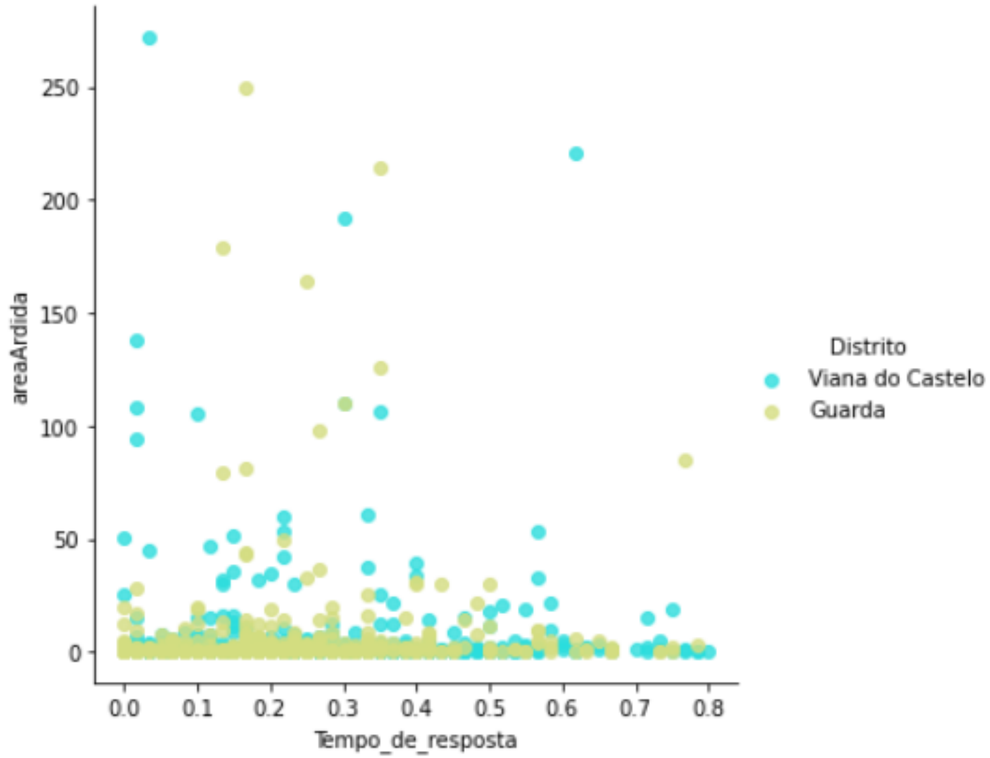
Inferences

Question 2:

- Is there a relation between the Burned Area and the time to Response that causes more or less impact?

IThe **covariance** between the two variables is almost **0 (1.85247199e-01)** suggesting the **variables are independent** as we expect.

We can see that the **two variables are not correlated** (value appr. 0) and that the **correlation is 0.066**.
This suggests a **no level of correlation**

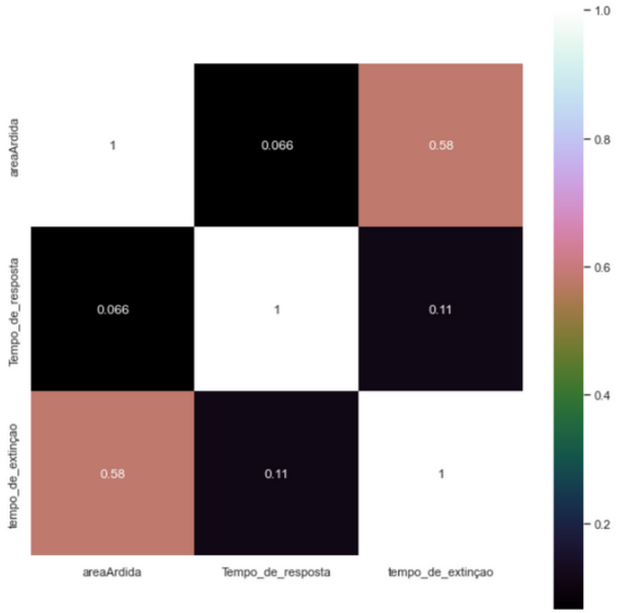


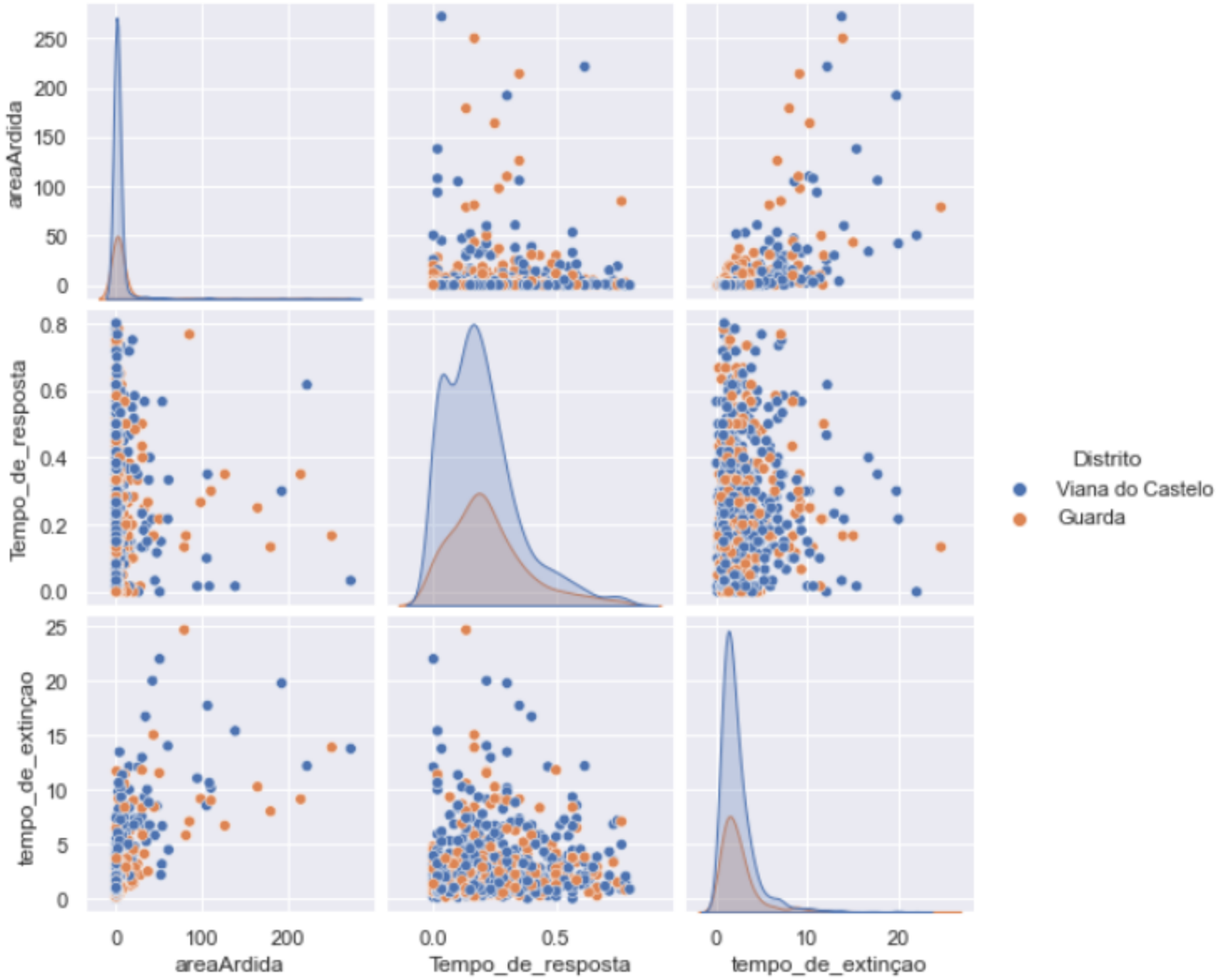
Inferences - Apendix

Significance Estimation

Question 2

Is there a relation between the Burned Area and the time to Response that causes more or less impact?





Conclusions

Conclusion 1

Guarda and Viana do Castelo have 18.9 % and 14.7 % of burned area out of a total of 68 acres in 2015.

Guarda presents a mean value of burned area per fire around 6 acres and Viana do Castelo around 3 acres.

They are apparently significantly different.
It seems that Guarda tends to have more burned area per fire than Viana do Castelo but we don't know the reasons for that. And for that we cannot state for causality.

Conclusion 2

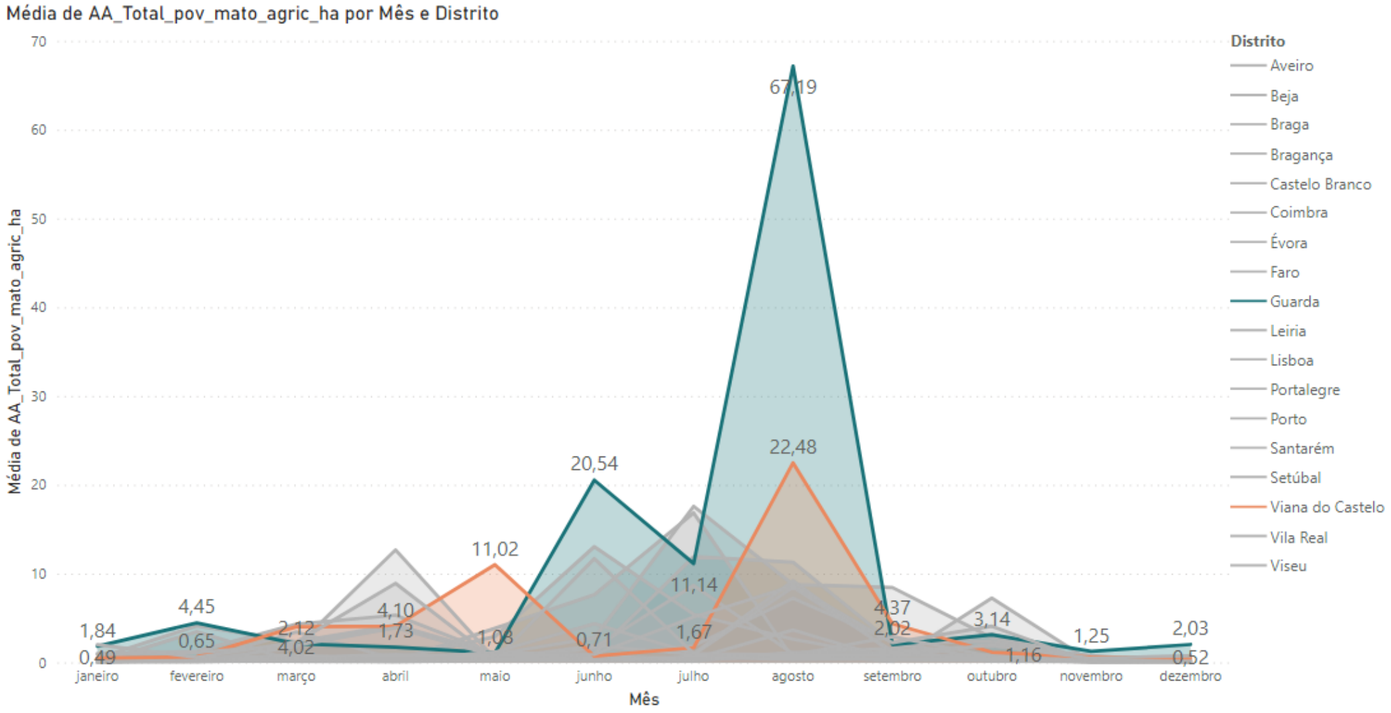
When studyng the relation between Burned Area and Time to response to a fire, we didn't found a significant relation among them.

Thus, is **not clear that a change on the Time to Response behavior will affect the impact of the forest fires relating the burned area**

Look back and ahead

More variables

For example, **Anual Sazonality, temperature values, wind.**

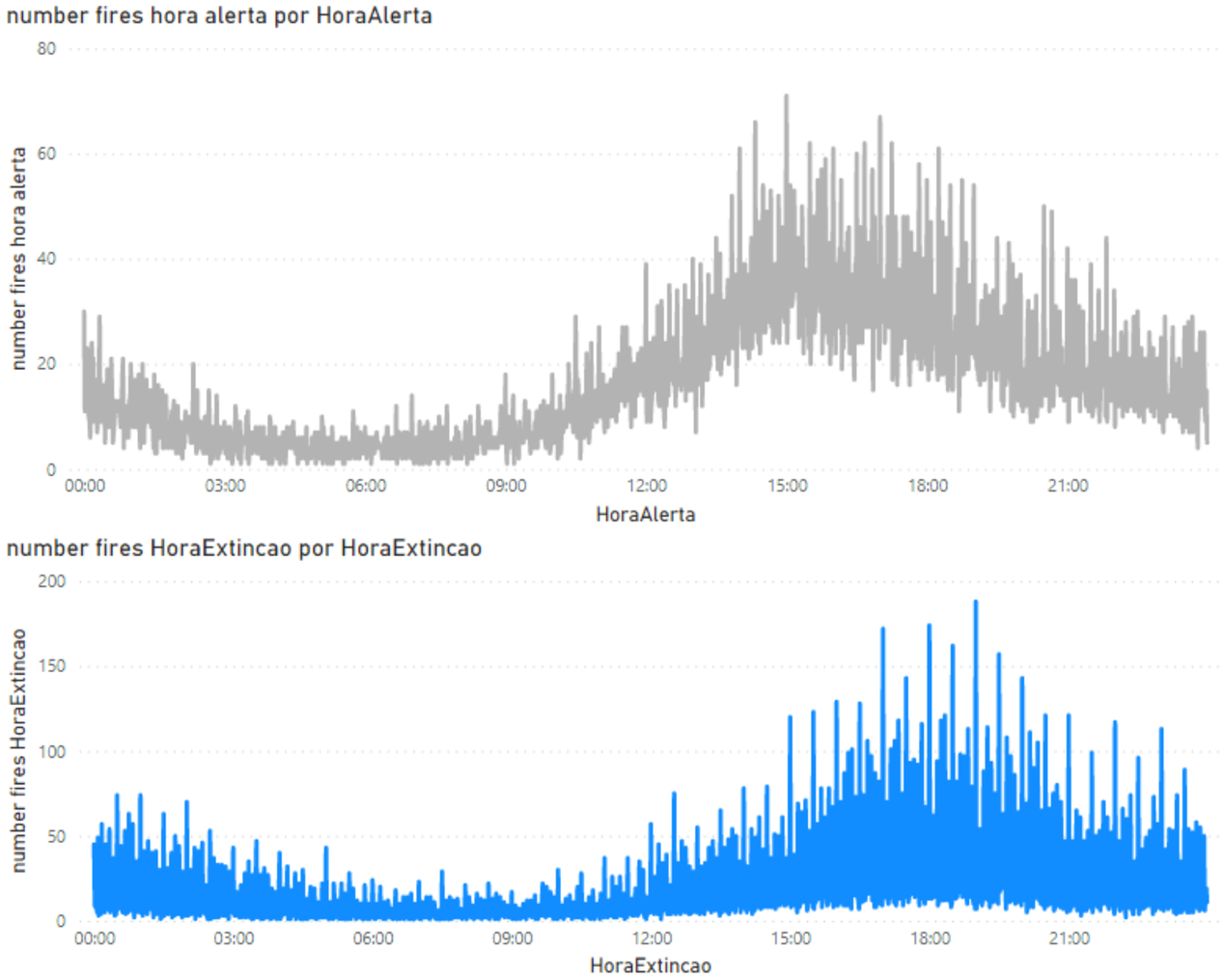


Mês	AA_Total_pov_mato_agric_ha	%GT AA_Total_pov_mato_agric_ha
agosto	30.374,51	44,45%
julho	11.146,41	16,31%
abril	6.832,48	10,00%
junho	5.201,50	7,61%
março	4.919,29	7,20%
setembro	4.475,23	6,55%
maio	2.846,22	4,17%
outubro	1.579,25	2,31%
fevereiro	534,31	0,78%
janeiro	207,35	0,30%
dezembro	146,51	0,21%
novembro	68,54	0,10%
Total	**68.331,59**	**100,00%**

Look back and ahead

By Hours

For example, **Daily Sazonality, temperature values**, **wind, humidity** related to the daily time of occurence



Look back and ahead

Other measures and locations

Measures like Fires frequency and Burned area could be combined to define a new measure for the fire impact

Other districts or other locations level should be inspected for evaluation (e.g. Concelhos)

Meet the Team

HUGO NOGUEIRA

FÁBIO FERNANDES